SOCCER VIDEO EVENT DETECTION VIA COLLABORATIVE TEXTUAL, AURAL AND VISUAL ANALYSIS

ALFIAN ABDUL HALIN

UNIVERSITI SAINS MALAYSIA
2011

SOCCER VIDEO EVENT DETECTION VIA COLLABORATIVE TEXTUAL, AURAL AND VISUAL ANALYSIS

by

ALFIAN ABDUL HALIN

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

October 2011

ACKNOWLEDGEMENTS

Alhamdulillaahi Rabbil 'Aalameen (All Praise Due to the One and Only God, Lord of All Creation). I thank Allaah, God the Almighty for providing me strength, guidance and patience to finally complete my PhD thesis. I pray to Him for success in this life and the hereafter.

Gratitude beyond words to my mother Sharifah Zaleha binti Syed Hassan, and my father Abdul Halin bin Hamid, for believing in me throughout this '*ordeal*' of a degree. Your continuous love, support and assistance, not only during my studies, but also throughout my life, can never be repaid. To my brother Izhal Abdul Halin (and wife Kak Farawahida Abdul Aziz), sister Afiza Abdul Halin (and husband Muhammad Jorge Arenas Abdullah), Mak Chaq and Kak Chik... Thank you for always being there when I needed you.

My deepest appreciation to my supervisor, Professor Dr. Mandava Rajeswari who has been a constant beacon throughout my PhD. Truly, I would have been in utter lost without you. To Dr. Dhanesh Ramachandram, thank you for your valuable insights. I would forever remember our chats about Canada and your undying love for photography.

Thank you to my lab mates Noridayu, Anusha, Osama, Adel, Jawarneh, Mozaher, Ehsan and Mogana. I have learned a great deal from all of you. Thanks for the support during the ups and downs of our 'battles' during this glorious process of a research degree.

All my friends and colleagues at UPM, you have provided comforting words during time of need. I am forever grateful for all the tips we shared. Thank you Mas

Rina, Nabil, Fadhlina, Azri, Noris, Noorsalwati, Amir, Kak Raja Azlina, Kak Salmi Baharom and also others.

My sincere thanks to my *Fishing Crew* of Azrul Hazri, Akhmal, Cikgu Othman, Malek and Sifu Mohet, for introducing me to the oceans of Yan, Kedah. Our Spanish Mackerel trips have been a stress reliever and I wait in anticipation for the next sea-sickness sessions. Many thanks also to Raja Azhar Raja Othman. I am already missing our FIFA-11 and HD movie sessions on your 32-inch Toshiba Regza flat screen LCD TV. Special thanks to friends, who constantly reminded me about life outside the lab. My gratitudes to Abu Marwan, Mohammed Azmi Al-Betar, Mu'awiya, Khalid, Asef, Ra'ed, Fazilah, Rashidi, Azleena, Aisyah, Bullert and Johnny Salleh.

To Dr. Sakrapee (Paul) Paisitkriangkrai, my ex-desk-neighbor at NICTA (*although only for* 2*-months*). I highly appreciate all the knowledge you've given me through our email and chat sessions. All the stuff about feature vectors, supervised and un-supervised learning and dimensionality reduction, have helped me a great deal in understanding the basics of Machine Learning. Although I haven't used them in my thesis, the knowledge you imparted will definitely be useful in my years to come as a researcher and as an academician.

Gratitude to my long-lost (and literally "*long*" as in 1.93-meters tall) friend Dr. Atheer Abas Matroud. Come to think of it, you were my first "PhD supervisor"! I am forever grateful for all the advice you've given me since 2007 until today. Owh! And not to mention all the LATEX tips and tricks, especially about using tables and charts! Although I believe I will never code any LATEX charts, thank you for letting me know that there are people out there who actually draw using words and numbers!!!

Thanks to the "*Hudson Street Boys*", Anang Hudaya and Cheah Wai Shiang. The

both of you inspire greatness through perseverance and hard work. May our paths cross again so that the Hudson Street days can be relived.

Last but definitely not least, thank you to my wife Nor Shazwan Mohamed. I can't think of anyone else with such patience, tolerance and love, and whose inner strength has propelled me past the finish line. Thank you my dearest for the support, encouragement and sacrifices. And thank you also, for going through approximately 18-months of pregnancy and countless hours of labor pain to give me (with God's blessings and permission) our beautiful daughter Nadiah Farhanah and handsome son Nadeem Farhan. My love to you all!!!

LIST OF PUBLICATIONS

Accepted Conference Proceedings and Colloquiums

1. Halin A. A., Rajeswari. M, Ramachandram. D., *Multimodal Video Analysis: A Review*, Computer Science Postgraduate Colloquium, 25-26 June 2007, Universiti Sains Malaysia.(Penang,Malaysia);

2. Halin, A. A., Rajeswari, M. and Ramachandram, D., *Overlaid text recognition for matching soccer-concept keywords*, 5th International Conference on Computer Graphics, Imaging and Visualisation, pp. 235-241. (Penang, Malaysia);

3. Halin, A. A., Rajeswari, M. and Ramachandram, D., *Automatic overlaid text detection, extraction and recognition for high level event/concept identification in soccer videos*, International Conference on Computer and Electrical Engineering, pp. 587-592. (Phuket, Thailand);

4. Halin A. A., M. Rajeswari, *Shot View Classification using HSV Color and Object Size*, in Computer Science Postgraduate Colloquium, 16-17 June 2009, Jerejak Resort and Spa. (Penang, Malaysia);

5. Halin, A., Rajeswari, M. and Ramachandram, D., *Shot view classification for playfield-based sports video*, IEEE International Conference on Signal and Image Processing Applications, pp. 410-414. (Kuala Lumpur, Malaysia).

Accepted Journal Papers

1. Halin, A. A., Rajeswari, M., Abbasnejad, E. M., *Soccer Event Detection via Collaborative Multimodal Feature Analysis and Candidate Ranking*, International

Arab Journal of Information Technology (Cited by: ISI, Impact Factor: *0.065*) - *Publication date to be later determined*;

2. Abbasnejad, M. E., Halin, A. A., Manshor, N., Rajeswari, M., *Automatic Image Annotation using Mixtures of the Exponential Family*, Journal of Convergence Information Technology (Cited by: SCOPUS, Engineering Village, IET InspecDirect, DBLP, DOI, Google Scholar, ProQuest, Scirus, Ultich's and EBSCO) - *Publication date to be later determined*.

Submitted

1. Halin, A. A., Rajeswari, M., Abbasnejad, E. M., *Goal Event Detection in Soccer Videos via Collaborative Multimodal analysis*, Pertanika Journal of Science and Technology (Cited by: Scopus and EBSCO) - Submitted August 2011 (*Has entered the second stage of the reviewing process*).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AGPR	Approximated non-event class representation for goals, penalties and red cards
ANECR	Approximated non-event class representation
ASR	Automatic speech recognition
AVI	Audio video interleaved
AYC	Approximated non-event class representation for yellow cards
BN	Bayesian network
CBIR	Content-based image retrieval
CC	Closed-captions
CCA	Connected components analysis
CLT	Central limit theorem
CT	Cinematic template
DBN	Dynamic Bayesian network
DVD	Digital versatile disk/Digital video disk
ESPN	Entertainment and Sports Programming Network
EUEFA	Union Of European Football Associations
FCM	Fuzzy c-means

FCM-HR	Fuzzy c-means hueristics rules
FIFA	Federation Internationale de Football Association
FN	False negative
FP	False positive
fps	Frames per second
FSM	Finite state machine
gb	Game break
HMM	Hidden Markov model
HSV	Hue, Saturation and Value
Hz	Hertz
I.I.D.	Independently and identically distributed
JPEG	Joint photographics expert group
LE	Log-energy
LLF	Low-level feature
LPC	Linear predictive coefficients
MAX/max	Maximum
MBM	Minute-by-minute
MFCC	Mel-frequency cepstral coefficients
MIN/min	Minimum

MLF	Mid-level feature
MP3	MPEG Layer-3
MPEG	Motion pictures expert group
MPS	Most prominent signature
OCR	Optical character recognition
QBIC	Query by image content
RGB	Red, Green and Blue
SBD	Shot boundary detection
SEE	Special editing effects
SMR	Slow motion replay
SSA	Semi-supervised approach
SSC	Semantic shot classification/class
SVM	Support Vector Machine
TN	True negative
TP	True positive
VOB	Video object
VOD	Video on demand
vs.	Versus
ZCR	Zero crossing rate

PENGESANAN ADEGAN VIDEO BOLA SEPAK MELALUI ANALISA KOLABORATIF TEKSTUAL, AURAL DAN VISUAL

ABSTRAK

Pengesanan adegan bola sepak berkait dengan mengenalpasti bahagian-bahagian menarik di dalam video bola sepak melalui analisa audio/visual. Ini membolehkan penghasilan indeks peringkat-tinggi secara automatik yang megelakkan anotasi manual berskala-besar dan memudahkan dapatan semula berasaskan semantik. Tesis ini mencadangkan dua rangka kerja pengesanan adegan melalui analsia kolaboratif tekstual, aural dan visual. Rangka-rangka kerja ini berkongsi satu komponen permulaan yang menggunakan sumber tekstual luaran, iaitu laporan per-minit dari penyiar sukan, bagi memusatkan secara tepat seksyen video beradegan yang dikehendaki. Rangka kerja pertama mengenalpasti anggaran permulaan bahaigan beradegan melalui analisa tenaga audio. Ciri semantik visual kemudian diamati bagi memperhalus bahagian beradegan yang dikesan. Rangka kerja kedua mengimplementasikan prosidur penyusunan berpangkat, di mana ciri semantik visual pada mulanya dianalisa bagi menjana senarai pendek calon-calon beradegan. Ini diikuti analisa aural atau visual di mana bahagian beradegan sebenar diberi pangkat teratas di dalam senarai pendek masing-masing. Kedua rangka kerja ini menggunakan ciri audio/visual yang mudah, dan ini merupakan kelebihan utama berbanding rangka-rangka kerja terdahulu. Data manual terlabel juga tidak diperlukan kerana pertimbangan audio/visual adalah berasaskan klasifikasi semantik visual automatik dan pengiraan ciri aural peringkat-rendah. Keputusan penilaian terhadap dataset video yang besar adalah memberangsangkan bagi pengesanan gol, penalti, kad kuning, kad merah dan penggantian.

SOCCER VIDEO EVENT DETECTION VIA COLLABORATIVE TEXTUAL, AURAL AND VISUAL ANALYSIS

ABSTRACT

Soccer event detection deals with identifying interesting segments in soccer video via audio/visual content analysis. This task enables automatic high-level index creation, which circumvents large-scale manual annotation and facilitates semantic-based retrieval. This thesis proposes two frameworks for event detection through collaborative analysis of textual, aural and visual features. The frameworks share a common initial component where both utilize an external textual resource, which is the minute-by-minute (MBM) reports from sports broadcasters, to accurately localize sections of video containing the desired events. The first framework identifies an initial estimate of an eventful segment via audio energy analysis. Visual semantic features are then observed to further refine the detected eventful segment. The second framework implements a ranking procedure where semantic visual features are firstly analyzed to generate a shortlist of candidates. This is followed by aural or visual analysis to rank the actual eventful candidate top-most within the respective shortlist. Both frameworks rely on uncomplicated audio/visual feature sets, which is the main advantage compared to previously proposed works. Furthermore, manually labeled data are not needed since audio/visual considerations are based on automatically classified semantic visual features and low-level aural calculations. Evaluation made over a large video dataset shows promising results for goal, penalty, yellow card, red card and substitution events detection.

CHAPTER 1

INTRODUCTION

1.1 Background

Catalyzed by rapidly advancing technology, the amounts of digital video being produced, archived and transmitted are reaching colossal proportions. Recording huge volumes is now feasible through relatively cheap devices such as digital camcorders (Rea, 2005; Smeaton, 2007). Personal archives have explosively grown with storage capacities in the region of gigabytes and terabytes. Compression schemes such as MPEG-2, DiVX and XVid are enabling high-quality video to be reduced to relatively small file sizes, avoiding excessive amounts of physical storage. Major improvements in bandwidth have also allowed seamless streaming and sharing of video over the world wide web. This proliferation, has led to information overload where efficient and timely access to video content is almost impossible (Elmagarmid et al., 1997). Production, generation and archival technology apparently have not been in tandem with the development of insightful tools for video data organization ad management. This can potentially render the data useless unless automated (or semi-automated) techniques are developed for proper annotation and indexing.

This need has spurred research interest where video annotation and indexing systems that extend Content-based Image Retrieval (CBIR) techniques began to emerge.

Notable works are such as Flickner et al. (1995) and Lee and Smeaton (2002) with the **QBIC** and **Fischlar** systems, respectively. However, since the CBIR paradigm revolves around low-level features , it is unsuitable for video's multidimensional semantic content. Therefore, intuitive approaches need to be considered where algorithms and/or systems are developed to understand video content at the semantic level (Rea, 2005).

1.2 Sports Video Analysis

Sports video are widespread and enjoyed by a global audience. The expanding viewership of current *video-on-demand* (VOD) services heavily cater for sports footage (Rea, 2005; Ren, 2008). Sports video archives are also burgeoning where the BBC motion library alone has a stockpile of over $20,000$-hours of multiple genre sports clips (BBC-Motion, 2010). These factors emphasize the need for automatic analysis techniques for effective consumption of sports video data.

The pervasiveness and volume of sports video necessitates automatic annotation and indexing techniques. This can facilitate many tasks such as video summarization, abstraction and retrieval applications. Since the effectiveness of a system strongly relies on the chosen index type, it is imperative to choose indices that best describes the video domain under consideration. For sports, this translates to event-based indexing since events are easily remembered and can often describe the nature of sporting matchups (Tjondronegoro et al., 2008; Ren, 2008; D'Orazio and Leo, 2010). Therefore, indexing at the event-level is considered the most practical and appropriate for sports video.

1.2.1 Sports Event Detection

Sports can naturally be decomposed into events. Interesting events such as *goals*, *yellow cards*, *red cards* and *off-sides* describe the essence of a soccer match similar to how *tries*, *penalties*, *drop-kicks* and *conversions* define a rugby game. Events are also easily recalled since they are more memorable and meaningful compared to other semantic concepts. Moreover, viewers' interest levels are normally retained only for short periods during event occurrences. Therefore, events are most suitable as semantic indices for organizing sports video.

Event detection however, is not clear-cut. It requires effective mapping of extractable audiovisual features to the respective high-level semantic concepts (i.e events). However, the existence of the *semantic gap* between machines and humans dictates the impracticality to develop generic approaches capable of detecting events across all sporting domains. Such a feat requires huge, if not infinite amounts of sporting knowledge (Ren, 2008; Rea, 2005). This problem however, can partially be alleviated by restricting the domain being addressed. Since different sports have different regulations, player dynamics, playfield geometry, and events, domain knowledge of a specific sport can be used to estimate the most accurate projections of features to events. Restricting the sporting domain not only allows for efficient use of domain knowledge, but also enables more specific event sets to be identified.

To focus the domain, this thesis concentrates on soccer videos. Soccer is one of the most popular sports on the planet and is a game with simple rules requiring minimal sporting equipment (FIFA, 2010; D'Orazio and Leo, 2010). Survey also indicates that huge demands for VOD are related to soccer footage in particular (Ren, 2008). Effec-

tive event-based indexing can therefore be useful to make soccer video content easily consumable by viewers/broadcasters. This thesis is motivated by this need, where investigation will be performed on video content analysis techniques for the detection of soccer events.

1.3 Significance of Work

The main significance of event detection is the ability to create more intuitive and memorable semantic indices, which leads to more effective annotations. Semantic level indexing and annotation can also facilitate the creation of useful applications such as summarization, abstraction and retrieval systems. The indexing paradigm is also shifted from low-level to high-level (semantic), which enables the specification of semantic text-based queries pertaining to the concepts being annotated. This avoids having to specify low-level feature combinations as done in the CBIR paradigm.

1.4 Problem Statement

Critical evaluation of the literature regarding previous approaches and frameworks for soccer event detection has led to the identification of the following issues:

1. *Varying and Complex Feature Considerations*: Many existing event detection approaches consider various semantic audio/visual feature classes. These are such as objects, motion parameters, playfield location, semantic shot classes and audio keyword classes. Most of these features require complex and multi-level/hierarchical detection processes, which can lead to poor classification ac-

curacy. Since effective event models necessitates reliable feature class identification, this can cause event models to fail. Moreover, many frameworks use a combination of varying semantic features to closely mimic the events under consideration. Although this approach can closely represent each event's audio/visual pattern, the modeling process relies on too many feature types whose detection accuracy is not perfect. An event model would benefit most with uncomplicated and limited feature considerations, especially those that can be reliably identified;

2. *Heuristic Rule-sets Applied to the Whole Video Duration*: Fully rule-based approaches require specifically defined heuristics to represent sequential and/or simultaneous feature occurrences during events. Rule modification and deletion is also simple, making their deployment to soccer video flexible. However, the main issue lies in the accuracy of such approaches due to the ambiguity in feature representations and the large search space the algorithms have to wade through. Soccer video is very noisy where non-events greatly outnumber the events. Moreover, different events can share similar feature representations, which causes certain algorithms to detect false alarms. Despite all being mentioned, rule-based approaches are still effective if the defined rule-sets are uncomplicated, and deployed in a less noisy search space;

3. *Limitations of Learning Algorithms*: The reliability of such approaches can still be questionable due to the lack of labeled positive (i.e. eventful) training examples. This is because sports video generally has huge content asymmetry where non-events greatly outnumber interesting events. Therefore, obtaining sufficient examples is difficult for reliable model construction (Ren, 2008). Supervised ap-

proaches also require extensive manual labor for labeling of training data, since no benchmark datasets are available such as that provided by the $TRECVID$ (Over et al., 2010) and $MediaMill$ (Snoek et al., 2005) initiatives for news video classification. Another issue is regarding model training time, which is a very lengthy process and normally requires tweaking of model parameters (Coldefy and Bouthemy, 2004). This can be arduous if many event models are to be generated. Although semi-supervised methods can be suggested as an alternative, such approaches require extensive effort for problem-model assumptions, determination of features and similarity functions, as well requiring that the data contain similar sample statistics.

1.5 Objectives

The primary aim is to develop an approach for detecting soccer events via multimodal (i.e. a combination between textual, aural and visual features) content analysis. The objectives can be further listed as:

- To investigate the use of suitable audio/visual feature types;
- To identify less complicated audio/visual feature combinations for event models consideration;
- To design and develop integrated components for effective feature processing and analysis;
- To develop a framework for event detection that involves minimal human intervention.

1.6 Scope

The scope of this thesis is defined in the following:

1. This thesis is concerned with soccer event detection, which can support useful applications such as summarization and retrieval. However, only event detection will be addressed as these applications are out of the scope of this work;

2. Event detection, with relation to indexing, is commonly an off-line process, where interesting segments are extracted after a match has ended (D'Orazio and Leo, 2010). Therefore, due to no real-time requirements, processing speed and time are not evaluated;

3. The dataset used in this work are recorded from various channels and broadcasters. It only consists of soccer footage where commercial intermissions, half-time analysis and match highlights are excluded. This is because segmentation of such content is a research area on its own (Kobla et al., 2000). Therefore, the work in this thesis will entirely focus on processing and analyzing soccer game footage.

1.7 Overview of Methodology

The methodology applied for the work in this thesis is multi-component-based, involving textual, visual and aural feature processing and analysis. The main task is to decompose a given match video into sub-segments, until ultimately (and ideally), only the desired eventful segment is identified. The initial components involve textual cues processing and utilization. For a given match video, it's minute-by-minute report is

utilized to obtain the name of events that have occurred along with their time-stamps. This information is used to localize event search to the eventful time-ranges derived from the time-stamps. After performing this localization, two approaches based on audio/visual analysis are investigated.

The first approach initially performs low-level aural feature extraction and analysis. The main idea is that soccer events causes a rise in audio energy. Therefore, identifying segments containing the strongest energy would highly likely indicate event occurrence. This is followed by refining the detected event segments via the visual observation of semantic shot classes.

The second approach on the other hand begins with mid-level visual analysis, where semantic shot classes are analyzed to decompose the localized segment into a shortlist of candidate segments. This is followed by a ranking process where each candidate within each shortlist are ranked based on their audio/visual feature signatures; where the top-ranked candidate is deemed to contain the actual event.

For both approaches, within each of the respective audio/visual components, established image and audio processing algorithms are applied. Detailed description of the frameworks are described in Chapters 5 and 6.

1.8 Contributions

The work in this thesis has led to three salient contributions. The first contribution pertains to semantic visual feature extraction (i.e. semantic shot classification) where an improved algorithm is proposed. The second and third contributions come in the

form of two frameworks for soccer event detection. Explanations for each contribution are as follows:

1. An improvement to an existing framework for semantic shot classification, which is detailed in Chapter 4:

    - The constraint of object-size is imposed, in addition to using playfield ratio as the determining factor for shot classification. Experimental results compared to an existing framework report significant improvements for *far* and *closeup*-view shot classification;

2. Two heuristic frameworks for the detection of *goal*, *penalty*, *yellow card*, *red card* and *substitution* events using textual and audiovisual content analysis (Chapters 5 and 6):

    - The first framework employs a generic approach for event detection where the aural channel plays the more important role. Thresholded audio energy peaks are determined within a localized event search space to identify segments most likely to contain the events. The visual features of semantic shot classes are then used to further refine the results to obtain the most accurate segments;

    - The second framework employs a more specific approach for event detection. A candidate extraction step is proposed where eventful segments are shortlisted using uncomplicated rules pertaining to semantic shot class transitions. Relevant segments are retained followed by a ranking strategy based on audio/visual event signatures. Specific ranking criterion is defined

for each event (termed as the *Most Prominent Signature* for an event candidate). The ranking process, which is done probabilistically or heuristically depending on the event, is different from existing approaches and provides the advantage of allowing actual event segments to still be accessible albeit ranking discrepancies;

- This framework also introduces the concept of *Approximated Non-event Class Representation* - ($ANECR$). The $ANECR$ of a specific match video represents the overall approximation of the audio feature probability distribution. This approximated representation is termed '*non-event representation*' as it exploits the extreme content asymmetry between non-events and events.

1.9 Thesis outline

Chapter 2 provides a critical review of related works pertaining to sports video event detection, with an emphasis on soccer. The chapter concludes with summarization of key points, which determines the direction taken by the work in this thesis. This is followed by **Chapter 3**, which describes the visual, aural and textual features being considered. **Chapter 4** presents an improved algorithm for semantic shot classification. Semantic shot classes are used as visual features in this work and this chapter explains how the algorithm is implemented. The proposed frameworks are then presented in **Chapters 5** and **6** followed by **Chapter 7**, which presents experimental results, comparison (for *goal* event detection) and discussions. **Chapter 8** concludes this thesis with remarks regarding limitations and possible future directions.

CHAPTER 2

LITERATURE REVIEW AND RELATED WORKS

2.1 Introduction

This chapter introduces the topic of event detection in sports video, with emphasis on soccer. Related literature are discussed, outlining the current state-of-the art in performing event detection. Specifically, the topic of feature selection and utilization is firstly addressed. This is followed by how the features are incorporated into event models, using specific event modeling approaches . The problems relating to the design of a comprehensive event detection framework is also addressed. This chapter concludes with a summary of the literature, as well as specifying the direction taken for the work in this thesis.

2.2 Sports Event Detection

Event detection is performed by firstly identifying the constituent audio/visual features that occur before, during or after an event. For example, a *goal* event is normally accompanied by rapid player movements during a far-view shot, followed by a rise in the crowd's and commentator's audio level, followed by a closeup-view shot, and ending with a replay showing a recap of the *goal*. These audio/visual observations are initially performed manually. After careful scrutiny, the observations are converted

into low or mid-level feature representations in order to allow computer processing and analysis. Event models are then constructed based on these features in order to detect similar events in other videos. The following sections will describe the common features used for soccer video, as well as the types of event models used to incorporate the features.

2.3 Feature Representation and Incorporation

The features normally extracted for event modeling problems are from the visual and aural channels. This is due to their direct availability within the video stream itself, therefore allowing established image and audio processing algorithms to be used for extraction. Basically, features can be extracted at the lower level (*i.e. low-level features - LLF*) and/or derived at the mid-level (*i.e. mid-level features - MLF*).

2.3.1 Visual Features

Visual LLFs include color histograms, edge densities, gray-level co-occurrence matrices, pixel intensity values etc. In sports video analysis, the use of visual features at the lower-level is less preferred due to the semantic gap. Even with the application of domain knowledge, calculations normally produce numeric representations at very high dimensions, which increases the complexity of event models. Therefore, most works rely on MLFs. Visual MLFs are semantic concepts, which are derived from specific combinations of LLFs. In soccer, the visual MLFs derived are related to the standard broadcasting practices of camera shooting style, and also pertaining to specific domain concepts such as playfield location the object categories involved. A summary of the

commonly used visual MLFs are provided in Tale 2.1.

2.3.1(a) Object Features

Important object categories in soccer either relate to the persons involved and also the soccer ball. Persons furthermore are divided into players and the referee. Successful identification or tracking of such objects is useful as visual cues for event occurrences.

Player identification is done through face detection and blob analysis. Faces are detected from video frames using algorithms such as proposed by Rowley et al. (1998); Jae-Ung et al. (2007); Snoek and Worring (2003); Sadlier and O'Connor (2005); Kolekar and Palaniappan (2008); Kolekar et al. (2009), which are suitable when the facial region is big. Identifying players at a distance requires a different treatment however, since the regions are small. This requires methods based on color differencing and background subtraction (Yasuo et al., 2006); or those that analyze static and dynamic pixel points by calculating pixel energy (D'Orazio, Leo, Spagnolo, Mazzeo, Mosca, Nitti and Distante, 2009). In Jia et al. (2009), players are detected through first performing background subtraction followed by boosting. Assfalg et al. (2003); and Bertini et al. (2004) on the other hand, detected players through blob analysis where potential player regions were identified matched with an elliptical template. Mostly, player identification is necessary for tracking and/or semantic shot labeling purposes.

The *referee* is also important as he normally appears during crucial events. Color projection profile analyses can be used to detect the distinct jersey color belonging to the referee. The main task is to identify the minimum bounding rectangle (MBR) in consecutive video frames to ascertain the referee's position. In Ekin et al. (2003), the

Table 2.1: Common Visual Mid-level Features

MLF Type	Derived Semantics	Method	Work
Object	Person (i.e. players and referee) and ball	**Person**: Face detection, blob detection and analysis, background detection and subtraction, color thresholding and minimum bounding rectangle (MBR) identification; **Ball**: Circularity, trajectory analysis, motion analysis, (small) blob detection and analysis and Kalman filtering, circular Hough-transform and color analysis	D'Orazio, Leo, Spagnolo, Nitti, Mosca and Distante (2009) Liu et al. (2009) Kolekar et al. (2009) Yasuo et al. (2006) Chung-Lin et al. (2006) Sadlier and O'Connor (2005) Bertini et al. (2004) Ekin et al. (2003) Snoek and Worring (2003) Assfalg et al. (2003)
Location	Playfield positions (e.g. goal area, penalty box, midfield, end zone etc.)	**All**: Edge detection, line detection, line orientation analysis, background detection and subtraction	Chung-Lin et al. (2006) Bertini et al. (2004) Assfalg et al. (2003) Ekin et al. (2003)
Motion	Camera parameters (e.g. fast pan, fast zoom etc.) and motion descriptors (e.g. rapid and lack of motion)	**Camera parameters**: Motion magnitude, motion vector displacement analysis, motion estimation	Hanjalic (2005) Leonardi et al. (2004) Coldefy and Bouthemy (2004) Bertini et al. (2004) Kongwah et al. (2003) Cabasson and Divakaran (2003) Kobla et al. (2000)
Semantic Shot Class	Camera shots (i.e. far view, closeup view, medium view and slow-motion replay)	**All**: Color analysis, background detection and ratio calculations and edge detection	Tjondronegoro and Chen (2010) Eldib et al. (2009) Abdul (2009) Tjondronegoro et al. (2008) Ren (2008) Changsheng, Wang, Lu and Zhang (2008) Kolekar and Palaniappan (2008) Kolekar et al. (2008) Min et al. (2006) Chung-Lin et al. (2006), Chung-Yuan et al. (2005) Zhou et al. (2005) Shu-Ching et al. (2004)

vertical and horizontal projection profiles of pixels were calculated to detect the referee's jersey and determine the encapsulating MBR. Chung-Lin et al. (2006) identified referee MBRs by assuming the referee wore black. Black pixel coordinates were then identified to obtain potential referee MBRs.

The *soccer ball* is useful as its trajectory indicates the occurrences of interesting events, especially pertaining to attack sequences. Earlier works used the circularity feature to determine ball positions across frames Gong et al. (1995) . This feature measures how close a particular shape resembles a circle. D'Orazio et al. (2002); D'Orazio, Leo, Spagnolo, Nitti, Mosca and Distante (2009) used a modified circle *Hough Transform* to detect the ball in video frames. Successful detection occurred especially when the ball has uniform color and is not occluded. Trajectory analysis based on camera motion can also be used. Xinguo et al. (2003) tracked ball trajectories in consecutive frames by firstly removing player blobs and noise. The trajectories were then estimated using the *Kalman* filter. The works of Bertini et al. (2004); Assfalg et al. (2002, 2003) captured motion features from the fixed main camera using three parameters namely the horizontal translations, vertical translations and isotropic scaling. They were concerned to identify rapid motion, which were assumed to be fast ball movement during events.

2.3.1(b) Playfield Position

The playfield position is a useful cue since important events cause progressions from one playfield zone to another. Mainly, playfield positions are determined through analyzing edges, lines and color related features. The penalty box and goal area were

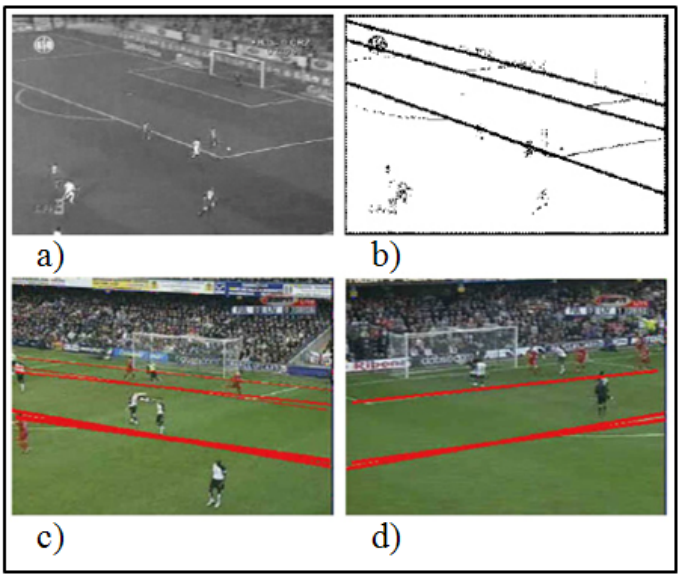

Figure 2.1: Penalty Box/Goal Post Detection: a-b are from Ekin et al. (2003) and c-d are from (Chung-Lin et al., 2006).

detected in Ekin et al. (2003); and Chung-Lin et al. (2006). Both use edge detection (Laplacian Edge Response and Roberts edge detectors, respectively) followed by the discovery of parallel lines via the Hough transform. Ekin et al. (2003) considered size, distance and parallelism constraints, while Chung-Lin et al. (2006) used angle ranges (i.e. 140° to 170° and 10° to 40°) and line tilt orientations. Some works identify playfield zones during the progression of matches. Assfalg et al. (2003); and Bertini et al. (2004) classify playfield zones with distinctive labels such as '*wide goal box, left*, *midfield, left* and *lower corner, left* (Figure 2.2). They firstly extracted the playfield region and lines using background subtraction and edge-detection algorithms, respectively. Next, numeric descriptors relating to the orientation and length of the lines as well as the shape of the extracted playfield are identified. The numeric descriptors were then used to feed a naïve Bayesian classifier to categorize each frame into the respective playfield zones, based on a set of 12-zones that were identified *a-priori*. Since events such as goals cause progressions from one zone to another, these labels are very useful in constructing event models.

Figure 2.2: The six zones of a playfield half.

2.3.1(c) Motion-based Features

Motion features can be useful at the lower-level. Hanjalic (2005); and Kobla et al. (2000) calculated motion magnitude along with other feature types from different channels to detect highlights in soccer and other sports programs. They argue that high motion segments normally correspond to video segment containing highlights. However, more intuitive features are still required at the mid-level. The motion MLFs relating to camera motion parameters such as panoramic and zoom factors, as well as

the estimation of fast or slow camera motion can be defined by examining the motion vectors and their displacements between frames. For instance, *fast pan*, *fast zoom* and *lack of motion* were inferred in (Leonardi et al., 2004) by calculating displacements between each *P*-frame in $MPEG$ bit streams. The value of *fast pan* and *fast zoom* for example, were determined by comparisons with predefined thresholds. When the thresholds are exceeded (i.e. 20 *for fast pan and* 0.002 *for fast zoom*), these camera motion parameters were considered to be present. Cabasson and Divakaran (2003) calculated average motion vector magnitude of *P*-frames to represent *high* and *low activity*. Kongwah et al. (2003) used motion direction and motion intensity calculated from motion vectors to detect *camera pan*. Their implementation accumulated motion vectors over three successive *I*/*P*-frames and were scaled to reflect the non-uniform camera pans with respect to the current playfield view position.

Motion estimation can also be performed to infer specific motion MLFs. Bertini et al. (2004) represented *player group acceleration* and *deceleration* by camera motion estimation. Motion magnitude and independent movements of pixel directions were clustered to identify the players' movements during a change in particular game action. Similarly, Coldefy and Bouthemy (2004) assumed dominant motion represented by a 2D-affine motion model corresponds directly with significant camera movement, which also signifies the occurrences of highlights and events.

2.3.1(d) Semantic Shot Classes

Video shots alone convey no semantic meaning. Therefore, semantic labels corresponding to the current camera shooting style such as *far*-views, *medium*-views, *closeup*

Figure 2.3: Left-to-right: *far* view, *medium* view and *closeup* view.

-views and *slow motion replays* can be useful. The main advantage of such shot classes over visual low-level features is that they simplify the event modeling process. Instead of considering real valued numerical descriptors, events can be modeled as how they might occur during broadcast. The process of assigning labels to shots can be termed as *semantic shot classification* (SSC). Examples of some semantic shot classes (SSC) are shwon in Figure 2.3.

The simplest method for SSC is to calcualte the playfield ratio (PR) within each video frame (Xu et al., 2001; Shu-Ching et al., 2004; Min et al., 2006; Chung-Lin et al., 2006). Each frame within a shot goes through color segmentation where the playfield region (grass) is firstly identified. This is done via either detecting the dominant color, or by detecting a color range pertaining to the color green. Next, calculation of the playfield to frame ratio is performed, where if the PR is over or a predetermined threshold, a particular view class is assigned. For example, if PR is above a 43%, then a frame is labeled as a *far*-view (Abdul, 2009). Such algorithms therefore, heavily rely on color features and also the various threshold considerations based on extensive observations.

More complicated SSC algorithms use a wider variety of features. The works in Kolekar and Palaniappan (2008); and Kolekar et al. (2008) for example use color,

motion and edge in a multilevel framework. In all, five specific detection steps are defined where at each level, specific threshold comparisons were made pertaining to dominant color and color ranges. Ten types of view classes were classified including *long*, *straight*, *corner*, *referee*, *players team A or B*, and *players gathering team A or B*. In Yu-Lin et al. (2003); Sun et al. (2003); and Yu-Lin et al. (2004), color, means and standard deviations of motion magnitudes, block-level motion vector angles and motion distribution were used in a two-level heuristic and SVM-based architecture to detect a wide range of shot classes. The authors claim that using very few classes is insufficient for proper event modeling. Therefore, more detailed shot class descriptions were introduced namely *far-view of whole field - goal post visible/not visible*, *far-view of half field - goal post visible/not visible*, *midrange-view active/passive - whole body of player visible/ not visible* and s*closeup-view of a player/referee/coach/goal keeper*.

Figure 2.4: Special editing effect of a logo transition.

2.3.1(e) Slow Motion Replays

Slow motion replays (SMR) are shots showing re-enactments of interesting events. Most algorithms classify shots as SMRs by detecting special editing effects (SEE) that sandwich the start and end boundaries of the replay (Tjondronegoro and Chen, 2010). Mostly, the SEEs are logo transitions of leagues or competitions. An example is adapted from Tong et al. (2008) and is shown in Fig. 2.4). Due to this, direct

feature representations pertaining to color, texture and motion as well as indirect properties such as frame sequence duration and shot duration are used to identify SEE appearances. Chung-Yuan et al. (2005) detected logo transitions by thresholding *hue* and *intensity* differences between pairs of consecutive frames. If the hue and intensity differences are 20 and 35, respectively, a logo transition is detected. Ren (2008) fused color histogram distance, shot duration, shot frequency, mean motion vector magnitude and grass area ratio into two Adaboost classifiers. The work in (Changsheng, Wang, Lu and Zhang, 2008) used a mean shift procedure to identify logo transitions and replay segments through spatio-temporal analysis of color, motion and texture features. More recently, Eldib et al. (2009) assumed the logo to be within the first 30-frames of a shot. They firstly performed image binarization, where candidate logo frames were identified if the white pixel count is more than 55%. This is followed by comparing the RGB mean of the candidate with the mean of preselected logo images. The logo is confirmed if it is between the pre-calculated threshold values.

2.3.2 Aural Features

Aural LLFs are commonly used as supporting cues, where their detection strengthens the hypothesis of an event. Hanjalic (2005) for example, used *audio energy* to complement motion features for the generation of an excitement curve. The peaks of this curve correspond to highlight instances such soccer goals, which exhibit significant increases in motion and audio energy. Snoek and Worring (2003) also used *audio energy* as a support to other visual MLFs for detecting interesting soccer events. Speech-band energy was used in (Sadlier and O'Connor, 2005) as the feature in a Support Vector Machine for eventful shots detection in various sports, which included soccer. They ar-

gue that shots exhibiting high audio measurements contain interesting events. Leonardi et al. (2004) calculated motion-based MLFs of camera parameters to obtain a candidate list for goal event segments. They finally used the *average loudness* difference between shot pairs to rank the candidates where higher difference values highly likely indicate that goals have occurred.

Another useful low-level feature is pitch. Tjondronegoro (2005) used sub-harmonic-to-harmonic ratio based pitch determination algorithm in (Xuejing, 2002) to obtain average pitch across audio clips. High-pitched segments normally correlate to excited speech, which is an important cue during event occurrences. Coldefy and Bouthemy (2004) estimated pitch readings within the $50Hz$-$500Hz$ band over 100-millisecond intervals. Used together with detected high motion measurements, soccer goal events were able to be identified.

2.3.2(a) Aural Mid-level Features

Besides being useful at the lower level, aural MLFs are also derived to represent certain semantics. Table 2.2 lists the various aural MLFs used across sports event detection literature. Inspection shows that the most widely used aural MLFs relate to sounds generated by the commentator(s) and crowd. This is unsurprising since during most sports broadcasts (especially soccer), interesting events are always accompanied by excitement from these two groups. The commentator's speech will normally be more rapid and excited during events, which is normally accompanied by crowd reactions via loud cheering. Therefore, such MLFs are good indicators of event occurrences. Mostly, energy, loudness, pitch strength, mel-frequency cepstral coefficients (MFCC),

log-energy and zero-crossing rate (ZCR) are used to derive these MLFs. In some cases, automatic speech recognition (ASR) is also applied (Hua-Yong and Tingting, 2009).

The referee's whistle is a cue mostly used for soccer broadcasts (e.g. Baillie and Jose (2003) and Tao et al. (2006)). The underlying principle is that certain events occur after the whistle is blown. For example, fouls and substitutions are preceded by a single short whistle. The full-time is signaled through double (or triple) long whistling. Detecting the referee's whistle can be accomplished by examining fluctuations in the audio zero-crossing rate.

Player-based MLFs commonly relate to ball-hits. This type of MLF is more appropriate for games involving the ball impacting a sporting instrument (e.g. a racket or bat), such as tennis (Zhenyan and Yap-Peng, 2005) and baseball (Rui et al., 2000; Ziyou et al., 2003). In baseball for example, ball-hits can be the main cue for events such as home-runs. For tennis, continuous ball-hits indicate that a rally is going on. Ball hits are usually detected using energy-based features.

2.3.3 Textual Feature Extraction

Textual features have received less attention compared to their visual and aural counterparts. This is because they are not directly available from the video itself. Utilizing the textual channel apparently can have great benefits since allows a video document to be as accessible as a textual document (Snoek, 2005).

Table 2.2: Aural Mid-level Features Variations

Work	Derived Semantics	LLF used for Derivation
Huang (2010)	Acclaim Silence	MFCC ZCR
Changsheng et al. (2009) Min et al. (2008)	Long/Double whistling Excited/Calm commentator speech Excited/Calm crowd sound	ZCR LPC SP
Ziyou (2005) Ziyou et al. (2005) Ziyou et al. (2003)	Applause Ball hit Crowd cheering Music Male/Female speech Speech with music	MFCC
Ballan et al. (2009)	Silence Speech Speech over crowd Crowd Excitement	MFCC Log energy
Hua-Yong and Tingting (2009)	Keywords: Goal and Penalty	ASR
Chih-Chieh and Chiou-Ting (2006)	Ball hit Crowd cheering Music Speech Speech with background music	ZCR Pitch MFCC
Tao et al. (2006)	Excited speech Referee's whistle	MFCC Energy Pitch
Zhenyan and Yap-Peng (2005)	Silence Speech Applause Ball Hitting	Volume ZCR MFCC
Jianguo et al. (2006)	Excited speech Referee's whistle	MFCC LPC Pitch
Baillie and Jose (2004) Baillie and Jose (2003)	Non-speech with low crowd sound Speech with low crowd sound Crowd cheering Crowd chanting Referee whistle Music within the stadium	MFCC Log-energy
Rui et al. (2000)	Exited speech Ball hits	Energy MFCC Information complexity Pitch